Usefulness of Factor and Cluster Analysis in Grouping Census Tracts

IRENE LEHR, Dr.P.H., HARLEY B. MESSINGER, M.D., Ph.D., and EDWARD S. ROGERS, M.D.

CENSUS TRACTS are useful demographic and statistical units of study. Within the larger circumscribed community, they are especially useful for determining subpopulation characteristics and subpopulations at risk to inadequate care, disease, and mortality. Officials participating in community health planning have expressed a strong interest in studies using census tract data. A recent report by the Study Group on the 1970 Census and Vital and Health Statistics suggests that the availability of such small area data "presents an excellent opportunity for those who are responsible for vital records systems to produce complementary data necessary for health planning, demography, and general social research" (*1*).

Census tracts provide useful basic statistical units for comparing vital rates only when both the number of events (the numerator) and the population at risk (the denominator) for each tract are adequate. To solve the problem of a numerator, the investigator can extend his observations to 3

Dr. Lehr was a research associate and Dr. Rogers was professor of public health and medical administration, School of Public Health, University of California, Berkeley, at the time of this study. Dr. Messinger is a fellow at the Institute of Health Research, Berkeley.

This study was supported in part by a Taxonomic Community Health Analysis grant HS-0012 from the Public Health Service and by General Research Support grant 5-S01-R-05441 from the National Institutes of Health. Computer services were supplied by Franck Oechsli, School of Public Health Statistical Services, and vital statistics data were supplied by the San Francisco Department of Public Health and Dr. Victor Eisner, clinical professor of maternal and child health, at the school of public health.

Tearsheet requests to Dr. Irene Lehr, 1786 Spruce St., Berkeley, Calif. 94709.

years centered on the census year. To solve the problem of a denominator, he is obliged to combine tracts. In the past, community health analysts working with census tract data and faced with these problems have either combined contiguous tracts judgmentally or selected a demographic measure and used it for grouping the tracts systematically (*2–5*). Such procedures risk combining tracts similar in one respect but different in others.

In this study, we have attempted to demonstrate the usefulness of an aggregate demographic structure for grouping tracts systematically so that meaningful rates of infrequent vital events can be calculated. (As we have used the term, an aggregate structure is a statistical construct in which a variety of variables for each unit is used for obtaining a basic structure of the whole.) An aggregate demographic structure is useful in community health studies only if results using the aggregate structure are compatible with results using nonaggregate data.

In earlier studies, it has been demonstrated that, by using an appropriate spectrum of demographic variables, factor analysis will explain the underlying structure of the given data (*6, 7*). Using such a derived factor structure and clustering techniques, census tracts with similar demographic structures can be systematically grouped together. Averaged factor profiles for groups of tracts can then be used for predicting infrequent events.

Prematurely born infants and infant deaths occur infrequently among most census tract populations (numbering only a few thousand people). Rates based on such early life events are sensitive health indicators that vary among population groups and because of environmental conditions (*8–10*). But, so far, researchers have found it difficult, if not impossible, to isolate the specific sociocultural components influencing early life mortality (*8, 10, 11*). Anderson has suggested that a reasonable approach would be to break the environmental components into fairly general ones and show how they operate at a given time and place (*10*). He stated that, in this way, the researcher would be able to determine the extent to which the various sociocultural components influence early life mortality.

The purpose of this study was to test the usefulness of factor analysis techniques for identifying the underlying demographic structure of census tract populations and then to use the derived structure for identifying similar tracts. Several early life health indicators such as lack of prenatal care, premature births, neonatal mortality, and infant mortality (events which are believed to have sociocultural causes and which tend to occur infrequently among census tract populations), were selected as the dependent variables for testing the utility of the derived demographic census tract constructs.

Procedure

The demographic data used in this study are from the 1960 Bureau of the Census figures for the San Francisco census tracts. The prenatal and infant data are taken by place of residence from the San Francisco Department of Public Health figures for 1959–61 for these same census tracts. The procedures used in analyzing these data follow.

Clustering variables. In an earlier factor analysis of 159 demographic variables for more than 800 census tracts in the six bay area counties (Alameda, Contra Costa, Marin, San Francisco, San Mateo, and Solano), it was observed that four factors accounted for most of the shared variance, or communality, of the variables. These factors, in turn, could be reproduced by 28 variables of the total set.

Using data for only the 121 census tracts in San Francisco and the reduced list of 28 salient demographic variables, preset key cluster analysis produced a satisfactory four-factor structure similar to the one for the entire bay area. In preset key cluster factoring (a special case of the general method of independent dimensional analysis developed by the late Prof. R. C. Tryon at the University of California, Berkeley), the researcher chooses as definers of each dimension the most collinear subset of variables that is also most nearly independent of the definers of other dimensions (*6*). The factors obtained by us from both bay area and San Francisco data are also similar to those obtained by others in earlier demographic analyses (*12–14*).

The four factors found to describe best the demographic characteristics of San Francisco census tracts were labeled (in order of their importance in accounting for covariation among tracts) socioeconomic status, family status, residential mobility, and ethnic status. These four factors, together with their coefficients and communalities, are given in table 1.

In this study of San Francisco census tracts the first factor, socioeconomic status, showed positive factor coefficients for the proportions higher education, occupation, and income variables, and negative coefficients for the proportions lower education and occupation measures. A crowding variable (proportion of housing units with 1.01 or more occupants per room) also had a negative coefficient (table 1). The second factor, family status, showed positive coefficients for such family measures as the proportions primary family households, couples with own household, one-unit structures, and owner-occupied housing units; it had negative coefficients for the proportions divorced white, widowed white, and persons over 65 years of age.

The third factor, residential mobility, had a positive factor coefficient for the proportion who had moved in 1958 to 1960; it showed negative coefficients for the proportions still in their 1955 residence, who moved in 1940 to 1953, couples with older children, and children under 18 living with both parents.

The correlation between the family status and residential mobility factors was higher in San Francisco (−0.732) than in the bay area as a whole (0.225). That these two factors should have such a high correlation in San Francisco and not in the bay area counties simply means that family status and residential mobility are related in San Francisco and not in the bay area as a whole. Because of the unfortunately high correlation in San Francisco between these two factors, stepwise rather than standard linear regression analysis was used for predicting the prenatal and infant rates.

The fourth factor, ethnic status, had positive coefficients for the proportions females in clerical or sales occupations and families with lower and middle income variables; it had negative coefficients for the proportions females in private household occupations, blacks, and races other than white or black. Structural purists will probably argue that the slightly higher coefficients for some of the female employment and middle income variables indicate that this is not really an ethnic factor. We agree that some of the variables in this factor are probably only indirect or even remote measures of ethnicity per se, but we also believe the characteristics of this factor, as a whole, are

Table 1. Factor analysis of San Francisco census tract demographic variables

Demographic variables	Factor coefficient	Total communality
Factor 1. Socioeconomic status:		
Proportion blue collar workers in employed male labor force [1]	0.977	0.975
Proportion males in professional occupations	.960	.930
Proportion completing college, 4 years or more [1]	.956	.943
Median school years completed [1]	.908	.853
Proportion completing elementary school, 5 to 7 years [1]	−.898	.838
Proportion housing units with 1.01 or more occupants per room	−.820	.778
Proportion blue collar workers in employed female labor force	−.803	.908
Proportion families with income $10,000 and over	.707	.837
Factor 2. Family status:		
Proportion primary family households [1]	.939	.919
Proportion married couples with own household [1]	.921	.912
Proportion units in 1-unit structures	.858	.818
Proportion enrolled in school	.845	.855
Proportion owner-occupied housing units	.824	.822
Proportion divorced, white	−.822	.795
Proportion 65 years old and over [1]	−.778	.679
Proportion widowed, white [1]	−.742	.604
Factor 3. Residential mobility:		
Proportion units moved into by present occupants in 1958 to 1960 [1]	.966	.944
Proportion still in 1955 residence [1]	−.959	.937
Proportion units moved into by present occupants in 1940 to 1953 [1]	−.904	.846
Proportion married couples with older children [1]	−.828	.812
Proportion children under 18 living with both parents	−.715	.595
Factor 4. Ethnic status:		
Proportion females in clerical occupations [1]	.815	.768
Proportion families with income $8,000 to $9,999 [1]	.789	.660
Proportion females in private household occupations [1]	−.621	.530
Proportion families with income $6,000 to $7,999 [1]	.584	.470
Proportion black	−.562	.590
Proportion females in sales occupations	.502	.391
Proportion other races (other than white or black)	−.348	.264

[1] Selected definer variables for each dimension.

Table 2. Factor scores for each census tract and mean factor scores for each core group of tracts [1]

Group and tract	Factor score			
	Factor 1	Factor 2	Factor 3	Factor 4
Group 1	42.0	33.2	35.0	35.3
A22	44.3	32.9	31.3	41.9
J9	47.8	32.3	33.7	33.0
J10	38.0	39.7	37.1	32.0
K2	38.1	27.8	38.0	34.3
Group 2	33.8	41.8	38.4	46.6
N8	34.1	40.5	39.6	42.5
N9	33.4	43.0	37.2	50.6
Group 3	39.2	44.3	58.1	40.2
A6	37.8	45.4	56.8	39.4
A13	45.6	46.6	58.8	42.1
A14	34.1	48.4	69.9	37.0
A15	42.6	40.7	60.8	41.9
A16	32.0	45.4	54.0	46.6
J3	43.1	39.1	48.5	34.3
Group 4	42.7	52.8	49.5	51.4
A1	48.9	55.9	50.5	46.9
A5	42.5	47.6	49.8	46.2
L1	42.3	59.2	50.5	44.8
L2	33.4	55.0	48.3	50.2
L3	37.6	56.6	52.2	48.9
M3	36.8	54.8	51.3	48.1
N3	41.4	47.3	46.0	48.5
N7	44.2	46.4	45.4	54.0
N10	41.7	50.0	45.8	54.4
N11	47.3	54.9	51.9	55.7
N12	45.0	55.8	51.0	55.8
N14	46.1	51.9	49.2	59.8
Group 5	37.6	61.9	57.8	51.7
L4	36.4	63.7	63.1	55.4
M1	36.4	61.0	55.2	53.3
M2	37.8	58.4	57.7	51.2
M4	40.0	59.0	53.7	48.7
Group 6	51.2	30.6	37.7	50.4
A17	51.8	34.5	42.8	52.4
A20	51.8	31.2	36.7	52.1
A21	52.4	22.3	36.1	50.1
A23	47.1	30.8	34.4	43.8
J1	52.7	34.4	38.4	53.4
Group 7	38.2	48.3	39.3	31.6
J6	42.9	48.5	42.9	30.4
J7	43.4	49.1	39.3	32.4
J8	41.9	44.8	38.1	28.4
J12	43.0	47.9	36.7	31.5
K3	30.0	44.2	42.8	33.2
K4	30.5	51.3	36.0	24.4
K6	37.8	55.9	41.4	35.1
N1	36.3	44.7	37.3	37.2
Group 8	47.3	44.9	40.8	43.0
J2	51.4	46.1	46.2	39.5
J11	42.7	40.3	37.8	45.4
J13	48.7	45.5	36.2	41.0
J14	50.7	48.3	42.2	47.1
J15	50.0	46.0	40.0	44.8
J16	47.9	47.9	42.4	37.1
J17	45.2	42.3	40.0	44.5
N2	41.4	42.8	41.9	44.5
Group 9	58.6	58.8	67.8	57.4
O3	62.5	57.8	60.9	55.6
O6	59.4	57.8	66.4	58.1
P2	56.9	59.9	70.6	59.7
Group 10	42.6	60.5	61.6	61.5
M5A	40.5	64.3	67.9	67.4
Group 10—Continued				
M5B	37.6	61.2	64.1	62.2
M6	40.1	59.2	61.0	61.2
M7	41.9	56.7	64.1	57.2
M8	44.0	62.6	65.2	61.6
M9	41.3	62.0	60.5	65.8
M11	41.8	65.3	64.0	65.8
N15	46.4	59.0	54.4	58.9
O9	45.5	56.8	58.6	53.9
Group 11	53.2	71.1	53.0	49.9
M10	45.7	70.3	48.1	50.5
R1	62.9	76.5	54.7	44.7
Group 12	55.8	62.0	61.4	65.7
N13	51.5	59.6	60.4	64.8
P1	55.0	51.9	59.9	64.6
Q1A	54.5	61.1	61.4	65.2
Q1B	56.3	67.2	68.0	67.9
Group 13	57.9	33.0	37.1	59.1
A10	54.4	37.4	38.5	63.2
A12	60.8	35.2	38.5	56.5
A18	61.6	28.7	35.8	59.5
A19	54.9	30.8	35.5	57.2
Group 14	64.6	41.1	44.4	43.9
A2	66.5	38.9	46.8	44.3
A11	61.6	43.8	44.8	49.0
B6	67.4	34.3	44.4	42.9
B9	62.6	45.7	44.7	44.2
B10	64.9	42.8	41.2	39.3
Group 15	64.8	47.8	51.6	35.4
B7	71.6	45.8	55.5	29.5
B8	65.8	52.0	48.2	36.7
J4	57.2	45.7	51.3	40.0
Group 16	57.1	48.4	48.3	55.6
A3	54.6	51.0	55.0	55.2
A4	56.2	52.1	50.8	52.5
A7	58.1	52.0	53.2	54.1
A8	61.7	41.6	47.3	57.0
A9	57.2	46.6	44.9	63.2
B1	61.4	44.4	48.4	55.6
B2	62.7	44.7	49.3	56.5
B3	61.5	45.3	50.8	53.7
B4	57.3	43.4	48.4	63.2
B5	60.6	45.6	46.7	50.8
D1	54.8	43.7	50.6	45.8
D2	56.3	46.2	51.0	50.6
E2	55.2	48.3	49.5	60.8
E3	55.7	45.6	49.9	51.3
G1	56.6	46.6	50.3	56.4
G2	55.5	45.2	53.1	58.6
G3	59.6	51.6	55.9	54.9
H1	51.3	48.7	49.2	51.7
H2	53.7	46.3	48.8	53.4
J5A	57.6	51.8	52.2	60.6
J5B	57.5	54.8	47.0	44.6
J18	48.0	47.4	45.1	57.5
J19	59.8	49.7	47.4	55.0
J20	56.7	51.2	43.2	51.0
N4	60.9	55.0	44.1	61.0
N6	49.0	51.5	46.5	58.7
O1	58.0	47.4	39.1	57.1
L2	33.4	55.0	48.3	50.2
Group 17	63.9	61.8	73.0	44.8
O4	62.0	54.8	65.9	45.8
O7	67.0	63.9	75.4	39.7

[1] Split census tracts omitted because data were not available were tracts L5B, N5A, N5B, O5A, O5B, O8A, O8B, P3A, P3B, and P3C. Census tracts omitted because their population characteristics were highly skewed were C1, G4, K1, and K5. Census tracts rejected from the object-typing because factor score patterns did not fit core group structures were E1 and L5A.

such that, without losing too much accuracy, we can apply the classic name of ethnic status to it.

A statistical factor is sometimes both difficult to visualize and name correctly because it usually represents more than one variable. Although it is a composite measure, it is usually given a simple name, preferably one which refers to its dominant variables. The name is more convenient to use, but it describes the factor only partially and inadequately. For this reason, it is important to think of a statistical factor more in terms of its variable structure than its name.

Clustering tracts. Using the previously mentioned structure of four factors composed of several variables each, an averaged score on each factor for each San Francisco census tract was obtained. Based on their factor-score patterns, the census tracts were then grouped by a computer object-typing program into 17 core groups of census tracts. These factor scores for each tract and each core group of tracts are given in table 2.

Tryon's condensation method was used for grouping tracts with similar factor scores. Using this method, the census tract factor scores were classified into high, H, medium, M, or low, L, tertiles on each of the four factors. Then, groups were formed of census tracts in all possible combinations of classes, that is, LLLL, LLLM, LLLH, LLML, LLHL, . . . , HHHH. Groups with at least two tracts were chosen as core groups; and using this typing procedure, we then assigned other tracts to the closest group on the basis of Euclidean distance. (For example, using the Tryon method, we found that the first core group of four tracts had a factor-score profile of 42.0, factor 1; 33.2, factor 2; 35.0, factor 3; and 35.3, factor 4. This core group of census tracts had factor scores below average on all four factors. The 12th core group of tracts, also four in number, had a factor-score profile of 55.8, factor 1; 62.0, factor 2; 61.4, factor 3; and 65.7, factor 4. This group of tracts was above average on all four factors.)

Groups of tracts which were too close were merged. Tracts too far from any core group can be rejected. We have described briefly the Tryon clustering procedure. Other procedures also can be used.

Deriving mean demographic factor scores. In the object-typing program, we also computed the mean on each of the four factors for each core group of tracts (table 2). The core means for each factor were used as the mean factor scores for each group of tracts.

Correlation coefficients between these mean factor scores (hereafter, these scores will simply be referred to as factors) based on data for the 17 core groups of tracts in San Francisco are given in table 3. Correlations between these factors and the rates per group of tracts are also given.

The computer programs employed in the preceding analyses are those of the BCTRY system of cluster and factor analysis, devised by Tryon and Bailey (*6, 15*).

Computing rates. Again using the core groups of similar census tracts, several rates relating to prenatal care or lack of it, prematurity, and mortality of infants were computed for each of the core groups of tracts. Statistics for 1959–61 were used to compute all rates. The four rates are (*a*) no-prenatal-care rate—no prenatal care or care only in the third trimester per 1,000 live births, (*b*) prematurity rate—births under 2,500 grams per 1,000 live births, (*c*) neonatal mortality rate—deaths from birth through 28 days per 1,000 live births, and (*d*) infant mortality rate—deaths under 1 year of age per 1,000 live births.

Predicting rates from the mean demographic factor scores. Finally, using the derived mean factor scores and the computed perinatal and infant rates for each of the core groups of tracts, stepwise linear regression analysis was used for predicting each of the rates from the four mean demographic factor scores.

Table 3. Correlation coefficients

Factors and rates	Factor			
	1	2	3	4
Factor 1, socioeconomic status.....	1.000	0.085	−0.289	0.181
Factor 2, family status.	.085	1.000	−.784	.348
Factor 3, residential mobility...........	.289	−.784	1.000	−.368
Factor 4, ethnic status.	.181	.348	−.368	1.000
No-prenatal-care rate..	−.654	−.525	.801	−.476
Prematurity rate......	−.210	−.720	.784	−.481
Neonatal mortality rate...............	−.109	−.454	.504	−.761
Infant mortality rate...	−.171	−.278	.379	−.833

NOTE: These correlations are based on data from 17 core groups. Thus, the factor data were the derived mean demographic factor scores for each group of tracts.

Results

No-prenatal-care rates. Stepwise linear regression analysis results, after the first step, indicated that the residential mobility factor explained 64 percent of the variation in core rates of women who had no care or had had inadequate prenatal care during pregnancy (table 4). The socioeconomic status factor, which entered at the second step of the analysis, boosted the explained variation to 84 percent; the multiple correlation (R) was 0.91. Although the coefficient of determination (R^2) indicated that 84 percent of the variance of the no-prenatal-care rate was explained by the two predictors, the F test showed a more significant figure for the second step than for the first—a dubious result. The ethnic status and family status factors, entered in the third and fourth steps, increased the explained variation to 87 percent. Because of possible collinearities among the four-factor demographic structure as used only for the data on San Francisco, we have taken a conservative position and have stopped with the first step of this regression analysis.

On this first and the following analyses, we did not seek the best set of predictors (as an alternative to stepwise regression) because this could have led to spurious results when the predictors are possibly collinear. The explaining of 87 percent of the variation, for example, is not likely to be reproducible if the regression equation were to be used on new data.

Infant mortality rates. Ethnic status and residential mobility were the first and second factors to enter this stepwise regression analysis on infant mortality rates. These first two factors explained 70 percent of the variation in infant deaths (table 4). Of these two, the ethnic status factor produced an R^2 of 0.69 and addition of the residential mobility factor brought the multiple R^2 to 0.70.

The family status and socioeconomic status factors entered at the third and fourth steps of this analysis. After the fourth step, the multiple R^2 was 0.71.

Again, because of possible collinearities among the demographic factors, we considered it better to use only the first two steps of this analysis.

Prematurity rates. Residential mobility and ethnic status, the first and second factors to enter this analysis, accounted for 66 percent of the variation in core group prematurity rates (table 4). Family status and socioeconomic status, the third and fourth factors to enter, increased the multiple R^2 to 68 percent.

As with the preceding results, we used only the first two steps of this stepwise regression analysis.

Neonatal mortality rates. The ethnic status and residential mobility factors entered at the first and second steps of this analysis of neonatal mortality rates. These two factors were slightly less useful for predicting the neonatal mortality rates than they were for predicting the prematurity rates or the infant mortality rates. Even so, the two accounted for 64 percent of the variation in neonatal mortality rates (table 4). After the first step, R^2 was 0.58.

The socioeconomic status and family status factors entered at the third and fourth steps of the analysis. After these last two steps, the multiple R^2 was still only 0.64.

Discussion

In this study, stepwise linear regression analyses and tests of significance showed that the demographic factors, either singly or in combination with a second factor, were highly significant predictors of each of the perinatal rates or infant

Table 4. Stepwise linear regression analysis results

Criterion variable	Multiple R	Multiple R^2	F value	Predictor variables	Regression coefficient
After step 1					
No-prenatal-care rate	0.801	0.642	[1] 26.863	Residential mobility	1.863
				Constant	147.388
After step 2					
Infant mortality rate	.837	.700	[1] 16.354	Ethnic status	−.484
				Residential mobility	.042
				Constant	50.834
Prematurity rate	.810	.657	[1] 13.390	Residential mobility	1.305
				Ethnic status factor	.496
				Constant	177.665
Neonatal mortality rate	.798	.636	[1] 12.246	Ethnic status	−.374
				Residential mobility	.121
				Constant	42.936

[1] 0.001 level of significance.

mortality rates. Within the San Francisco four-factor demographic structure, three of the factors were especially efficient for predicting the several health or mortality rates.

These results are impressive statistically, but do they confirm what we already know empirically? Do results from these analyses in which clustered variables and clustered census tract populations were used agree with observations of studies in which individual data and more modest methods were employed?

Ethnic status. Results of stepwise linear multiple regression showed that, of the four demographic factors, ethnic status was the most efficient for predicting the infant and the neonatal mortality rates. These results agree with national, State, and San Francisco figures, which indicate that the black infant is subject to excessive mortality (*16–18*). In San Francisco, both the neonatal and the infant mortality rates are higher among blacks than among whites, and rates for both blacks and whites are higher than those for the Chinese, Japanese, and all others (*18*). Many other analyses have also confirmed that infant deaths are higher among blacks than among whites (*19*).

Socioeconomic status. Of the four predictors, the socioeconomic status factor was inefficient in predicting the prematurity rate and the neonatal and infant mortality rates. This discovery is contrary to our general belief that prematurity and mortality, especially preventable deaths, are caused partly by socioeconomic circumstances. This belief may actually be true, but in conjunction with other conditions, these other conditions may dominate the socioeconomic circumstances to the point where the prematurity or death of infants is relatively independent of socioeconomic circumstances and much more dependent on other conditions.

Medical authorities have long associated an inadequate income with lack of prenatal care. They have also associated prematurity with socioeconomic status and lack of prenatal care, but Terris and Gold have suggested that this last conclusion is open to debate (*20*). In this study, residential mobility and probably socioeconomic status were highly efficient predictors of adequacy of prenatal care. These results support the idea that a lower socioeconomic status is associated with lack of prenatal care, but the results also indicate that residential mobility is even more importantly related to both lack of care and prematurity.

One can conclude from these analyses of data on San Francisco census tracts that premature births and neonatal and infant mortality rates are influenced more by ethnic status and residential mobility than by socioeconomic status.

Residential mobility. Stepwise linear regression analysis indicated that, among the four demographic predictors, the residential mobility factor was the most efficient predictor of both the lack of care and the prematurity rates. In fact, the residential mobility factor was apparently better than the socioeconomic status factor for predicting lack of care. The predictive power of the ethnic status factor, however, dominated that of the residential mobility factor for predicting neonatal and infant mortality.

Residential mobility and its associations with lack of prenatal care and prematurity has received only secondary attention. Terris and Gold, studying premature births of black infants in hospital wards, noted that prematurity was directly associated with both length of residence in New York City and a maternal history of premature births (*20*). These results might be true for New York City blacks but not be true for nonblacks or for blacks in other sections of the country— national figures for both whites and nonwhites indicate that prematurity rates tend to vary with size and place of residence (*21*).

Some national statistics lend indirect support to our discovery that residential mobility is associated with prematurity and lack of prenatal care. For example, natality figures of the United States show that young nonwhite women less than 20 years of age have higher birth rates than young white women of the same age; that young mothers, especially nonwhites, have more premature births; and that prematurely born infants account for a greater proportion of newborn deaths (*22*).

Other statistics are probably available which indicate that young families, especially nonwhites, are residentially more mobile than older white families; and that new young residents, especially young nonwhites, may either lack the funds necessary for medical services or, even with prepaid coverage, may be unaware or disregard the importance of early prenatal care in preventing premature or low weight infants.

Family status. Among the four predictors, the family status factor was inefficient for predicting any of the several rates because it was overshadowed by the other three factors. By itself, however, it could have explained half the variance of prematurity rates since the correlation of these two variables was —0.720. Because the correlation of residential mobility was higher (0.784), it took precedence. The correlation of family status and residential mobility was —0.784, so family status was no longer efficient once residential mobility was in the equation.

Clustering variables and tracts. Results from these regression analyses of demographic factors on rates for groups of census tracts indicate the usefulness of factor analysis techniques. Their use in obtaining demographic factors and in clustering tracts may be as productive as, or more productive than, the usual methods of accumulating and analyzing data for individual persons and of grouping census tracts.

A researcher into sociocultural problems usually works with a large number of variables. Using factor analysis, he can group the many characteristics into a reduced number of factors. Such factors, derived from the clustering of variables, represent the basic structure underlying the observed characteristics. Using factor analysis, the researcher can also determine the number of factors that are necessary to account for the greater portion of covariance among the total configuration of characteristics. Once the observed characteristics have been reduced and the necessary clusters of variables have been obtained, he can then use the derived factor structure for classifying different population groups.

Census tract studies of sociocultural problems sometimes suffer because the number of events per year is small. When events are few, the researcher can use averaged rates, that is, the summing of several years' events, but this procedure is not always satisfactory because the base population may change in size or character (thus we used only the 3-year period centered on the census year 1960). We suggest that researchers use derived factor scores for grouping tracts that have similar demographic structures. A mean rate per cluster of tracts can then be computed. This procedure of clustering tracts that have similar factor structures can be likened to the processes used for classifying plants or animals; both procedures are substantive as well as methodological. And in both procedures, the researcher is able to classify subjects without imputing causative dynamics to the classification structure.

The usefulness of factor analysis for reducing a large number of variables has been demonstrated many times. Through this study we merely wish to point out the usefulness of derived factor structures for grouping similar census tracts.

Although the primary purpose of this study was to test the utility of factor and cluster analysis techniques, we would be remiss if we did not discuss the practical uses of the results of these regression analyses. Although some of the regression results were expected, others were not. We should have expected the ethnic status factor to be a good predictor of infant and neonatal mortality rates. We were surprised, however, to discover that the residential mobility factor was such a good predictor of lack of prenatal care and prematurity. But, before these results can be used for program planning purposes, we must examine more carefully each of these factors and their component parts, that is, their variables.

For example, we must realize that the residential mobility factor includes three important variables (according to their factor coefficients) which are explicit measures of residential mobility. This factor includes two less important variables (proportion married couples with older children and proportion children under 18 living with both parents) which are probably less definite measures of residential mobility.

Also, the ethnic status factor includes variables with negative coefficients (proportion females in private household occupations, proportion black, and proportion of other races) which are more explicit measures of ethnicity. The ethnic status factor also includes variables with positive coefficients (proportion females in clerical occupations, proportion families with income from $8,000 to $9,999, proportion families with income from $6,000 to $7,999, and proportion females in sales occupations) which are less explicit measures of ethnicity.

Thus, to interpret these regression analysis results correctly, there must not only be an awareness of each variable in a factor but also an awareness of the special contribution each variable makes to the more general dimension. And, since the factor coefficient of a variable is simply its estimated correlation with a hypothetical score on that factor, one can easily determine the contribution of a variable to its factor by examining its factor coefficient.

Community health workers should be able to use the results of these regression analyses for planning. For example, by knowing that residential mobility is a good predictor of lack of prenatal care and prematurity and by determining which census tracts have a high score on the residential mobility factor and a high birth rate, they can then see that the health services needed by this particular high-risk group living in certain high-risk areas are readily available through appropriately trained personnel and accessible facilities.

It would be wise, as soon as the 1970 demographic data for San Francisco census tracts become available, to develop a factor structure based on the reduction and analysis of San Francisco data only. We would expect the 1970 factor structure to be similar to the 1960 one, but the newer structure would probably describe the population better because it would be based on San Francisco data only. The newer structure, therefore, would probably be more useful for clustering census tracts.

Using the 1970 factor structure and health or mortality rates averaged around the 1970 census for each tract, the various mean factor scores and rates for similar groups of tracts could again be computed. Then, by using 1970 mean factor scores and rates for clusters of tracts, the regression analysis results should be even more useful than the 1960 results for program planning.

Summary

Factor analysis was used for determining the salient demographic properties of San Francisco census tracts in 1960 and for deriving a four-factor demographic structure. This derived structure was then used for classifying almost all the 121 census tracts into one of 17 core groups with the aid of Tryon's condensation method. Mean demographic factor scores were computed for each of these tract clusters.

Perinatal and infant mortality rates (1959–61 averages) for each of the 17 core groups of tracts were computed. Stepwise multiple regression analysis indicated that, among the four predictors, the ethnic status factor was best for predicting the infant and the neonatal mortality rates; the residential mobility factor was best for predicting the no-prenatal-care and the prematurity rates. These results compare favorably with observations from the more traditional analyses of prematurity and infant mortality.

We suggest that factor structures, obtained from factor analysis of a larger number of variables and used to derive homogeneous census tract clusters, are useful adjuncts to the study of infrequent but complex disease or mortality events.

REFERENCES

(*1*) National Center for Health Statistics: The 1970 census and vital health statistics: a study group report of the public health conference on records and statistics. U.S. Government Printing Office, Washington, D.C., 1969.

(*2*) Faris, R. E. L., and Dunham, W. W.: Mental disorders in urban areas: an ecological study of schizophrenia and other psychoses. Hafner Publishing Company, Inc., New York, 1960.

(*3*) Chiazze, L., Jr., and Ciocco, A.: Intra-community variation in cancer incidence for Pittsburgh. Public Health Rep 82: 759–770, September 1967.

(*4*) Donabedian, A., Rosenfeld, L. S., and Southern, E. M.: Infant mortality and socioeconomic status in a metropolitan community. Public Health Rep 80: 1083–1094, December 1965.

(*5*) Bedger, J. E., Gelperin, A., and Jacobs, E. E.: Socioeconomic characteristics in relation to maternal and child health. Public Health Rep 81: 829–833, September 1966.

(*6*) Tryon, R. C., and Bailey, D. E.: Cluster analysis. McGraw-Hill Book Co., New York, 1970.

(*7*) Rogers, E. S., and Messinger, H. B.: Human ecology: toward a holistic method. Milbank Mem Fund Q 45: 25–42, January 1967.

(*8*) Richmond, J. B., and Weinberger, H. L.: Session II—program implications of new knowledge regarding the physical, intellectual, and emotional growth and development and the unmet needs of children and youth. Am J Public Health 60: 23–73, April 1970; pt. 2.

(*9*) Infant mortality—international comparisons. Stat Bull Metropol Life Ins Co 51: 3–4, February 1970.

(*10*) Anderson, O. W.: Infant mortality and social and cultural factors: historical trends and current patterns. *In* Patients, physicians and illness, edited by E. G. Jaco. Free Press, Glencoe, Ill., 1958, p. 11.

(*11*) Rennard, M.: Perinatal mortality: a review of 450 consecutive perinatal deaths. Am J Obstet Gynecol 104: 733, July 1, 1969.

(*12*) Bell, W.: The social areas of the San Francisco bay area. Am Sociol Rev 18: 39–47, February 1953.

(*13*) Shevky, E., and Bell, W.: Social area analysis: theory, illustrative application and computational procedures. Stanford University Press, Stanford, 1955.

(*14*) Tryon, R. C.: Identification of social areas by cluster analysis: a general method with an application to the San Francisco bay area. University of California Press, Berkeley and Los Angeles, 1955.

(*15*) Tryon, R. C., and Bailey, D. E.: The BCTRY computer system of cluster and factor analysis. Multivariate Behavioral Research 1: 95-111, January 1966.

(*16*) National Center for Health Statistics: Vital statistics of the United States, 1966. Vol. 2. Mortality, pt. A, sec. 2, Infant mortality. U.S. Government Printing Office, Washington, D.C., 1968.

(17) California State Department of Public Health: Vital statistics of California, 1965–1966–1967. Berkeley, 1969, p. 13.
(18) San Francisco City and County Public Health Department: Statistical report, 1960. San Francisco, p. 19.
(19) Chase, H. C.: White-nonwhite mortality differentials in the United States. Health, Education, and Welfare Indicators, June 1965, pp. 30-31.
(20) Terris, M., and Gold, E. M.: An epidemiological study of prematurity. 2. Relation to prenatal care, birth interval, residential history, and outcome of previous pregnancies. Am J Obstet Gynecol 103: 374, Feb. 1, 1969.
(21) National Center for Health Statistics: Vital statistics of the United States, 1966. Vol. 2. Mortality, pt. B, sec. 7, Geographic detail for mortality. U.S. Government Printing Office, Washington, D.C., 1968.
(22) National Center for Health Statistics: Natality statistics analysis, United States, 1962. U.S. Government Printing Office, Washington, D.C., 1964.

LEHR, IRENE (University of California, Berkeley), MESSINGER, HARLEY B., and ROGERS, EDWARD S.: *Usefulness of factor and cluster analysis in grouping census tracts. HSMHA Health Reports, Vol. 87, February 1972, pp. 155–163.*

In this study, the usefulness of factor and cluster analysis techniques for grouping census tracts was tested so that meaningful rates for infrequent vital events could be calculated. Factor analysis was used to determine the salient demographic properties of San Francisco census tracts in 1960 and to derive a four-factor demographic structure. This derived structure and Tryon's condensation clustering method were then used to classify almost all the 121 census tracts into one of 17 core groups. Mean demographic factor scores were computed for each of the census tract clusters.

Perinatal and infant mortality rates (1959–61 averages) for each of the 17 core groups of tracts were computed. Stepwise regression analysis indicated that, among the four predictors, the ethnic status factor was best for predicting the infant and neonatal mortality rates, and the residential mobility factor was best for predicting the no-prenatal-care and the prematurity rates. These results for the infant and neonatal mortality rates were similar to observations from the more traditional analyses; the results for the no-prenatal-care and the prematurity rates, however, were new discoveries. They are useful to community health workers responsible for health services planning.

We conclude from these analyses that factor structures, obtained from factor analysis of a large number of demographic variables and used to derive homogeneous census tract clusters, are useful adjuncts to the study of infrequent but complex health or mortality events.